The ISC Int'l Journal of Information Security

August 2019, Volume 11, Number 3 (pp. 79–89)

http://www.isecure-journal.org

Selected Paper at the ICCMIT'19 in Vienna, Austria

Enhancing Learning from Imbalanced Classes via Data Preprocessing: A Data-Driven Application in Metabolomics Data Mining☆

Ahmed BaniMustafa [1,*]

[1] *Computer Science Department, American University of Madaba, Kings Highway, Madaba, Jordan*

ARTICLE INFO.

Keywords:
Data Mining; Metabolomics; Cachexia; Preprocessing; Imbalanced Classes; Re-sampling; Data Reduction

Abstract

This paper presents a data mining application in metabolomics. It aims at building an enhanced machine learning classifier that can be used for diagnosing cachexia syndrome and identifying its involved biomarkers. To achieve this goal, a data-driven analysis is carried out using a public dataset consisting of 1H-NMR metabolite profile. This dataset suffers from the problem of imbalanced classes which is known to deteriorate the performance of classifiers. It also influences its validity and generalizablity. The classification models in this study were built using five machine learning algorithms known as PLS-DA, MLP, SVM, C4.5 and ID3. This model is built after carrying out a number of intensive data preprocessing procedures to tackle the problem of imbalanced classes and improve the performance of the constructed classifiers. These procedures involves applying data transformation, normalization, standardization, re-sampling and data reduction procedures using a number of variables importance scorers. The best performance was achieved by building an MLP model that was trained and tested using five-fold cross-validation using datasets that were re-sampled using SMOTE method and then reduced using SVM variable importance scorer. This model was successful in classifying samples with excellent accuracy and also in identifying the potential disease biomarkers. The results confirm the validity of metabolomics data mining for diagnosis of cachexia. It also emphasizes the importance of data preprocessing procedures such as sampling and data reduction for improving data mining results, particularly when data suffers from the problem of imbalanced classes.

1 Introduction

The muscle wasting syndrome or cachexia is a disease that is usually associated with cancer chemotherapy or radiotherapy. It may affect the lifestyle and life-expectancy of patients. Yet, it is difficult to observe, particularly in obese or overweight individuals as it may be independent of weight loss or changes in fat mass. Current diagnosis of cachexia involves muscle mass quantification and image analysis using X-Ray computed tomography (CT scan),

☆ The ICCMIT'19 program committee effort is highly acknowledged for reviewing this paper.

* Corresponding author.

Email address: banimustafa@gmail.com

ISSN: 2008-2045

Dual-energy X-ray Absorptiometry (DXA), and magnetic resonance imaging (MRI). These methods are expensive, time-consuming and expose patients to doses of radiation [1]. Hence, this research aims at using metabolomics data mining for the purpose of developing alternatives for the current diagnosis techniques.

Metabolomics is defined as: "the study of all low-molecular-weight chemicals that are involved in metabolism as an end product, intermediate or necessary chemicals" [2]. Metabolomics is usually performed using four main approaches: (1) Metabolite Profiling; (2) Metabolite Target Analysis; (3) True Metabolomics; and (4) Metabolic Fingerprinting [3, 4]. Data mining was used successfully in several scientific applications[5, 6]. The original study from which the dataset was acquired aimed at identifying and measuring all metabolites that are involved as biomarkers for cancer-related muscle loss syndrome in order to develop a simple, rapid and cheap test that can be used for screening the disease or for providing an indication for its likelihood [1]. Metabolomics data mining was applied successfully to health assessment, disease diagnosis and drugs monitoring [7–9]. It can be used to identify a number of metabolites that can be used as biomarkers for the cachexia syndrome. This would help to develop a rapid, cheap and robust diagnosis tool that can be safer and more accessible than the current diagnostic tools exposing the patient to radiation.

In this research, a data-driven and data mining based metabolomics diagnosis is carried out using four machine learning techniques known as Partial Least-Square Discriminant Analysis (PLS-DA), Multilayer Preceptor (MLP), Support Vector Machines (SVM) and Decision Trees (DT) using C4.5 and ID3 algorithms. In this research, we aim at studying the effect of various preprocessing procedures on learning from imbalanced classes as well at verifying, confirming and improving the results reported in earlier studies. Data preprocessing procedures were performed mainly for the purpose of tackling the issue of imbalanced classes which concerns the sample's distribution between control and patient groups [10, 11], in addition to improving the quality of the data and enhancing the validity of the results. It covers data transformation, standardization, normalization, reduction, and sampling. The data reduction was performed using: Information Gain (IG), Random Forests (RF) and SVM, while the sampling was performed using: Random Re-Sampling, Under-sampling, and SMOTE Over-sampling [12, 13]. The models were trained and then evaluated using five-fold cross-validation in order to ensure the validity of the results and assess the performance of the created model and also to be consistent with the evaluation methods that were used in the original study [1]. The champion models which are reported in this research both: (1) achieves the analytical objectives and (2) out-perform all the other generated models created in this research and the base model reported in the original study.

2 Materials and Methods

This section provides an overview of the dataset that was used in the analysis as well as the machine learning techniques that were applied and their evaluation mechanisms.

2.1 The Metabolomics Dataset

The data used in this paper is a publicly available dataset that represents a 1H-NMR metabolite profile. The profile consists of 77 samples that include a total of 63 identified and measured metabolites. About 47 samples were collected from patients with symptoms of cachexia, while the rest were collected from a control healthy group. Bio-fluid urine samples were collected from patients during their normal visits to cancer clinics randomly during the day. The distribution of both groups was similar in term of age and gender.

The dataset was acquired using a 1-Dimensional spectrum using the 1st increment of standard NOESY pulse NMR sequence at 600 MHz. Spiking was carried out using Human-Metabolome reference library. Reverse-phase HPLC was carried out on spiked samples based on Pico-Tag method. The metabolite profile was created to include all the identified and quantified metabolites. More details regarding the dataset acquisition and the design of the biological study are available in [1].

2.2 The Applied Data Mining Techniques

The data mining techniques applied in this research covers four supervised learning techniques including PLS-DA, MLP, SVM and Decision Trees using both C4.5 and ID3 implementation algorithms. The nature of the dataset and its size regarding the number of samples and attributes in addition to the analytical objectives influenced the choice of the applied machine learning techniques.

2.2.1 PLS-DA

Partial Least-Square Analysis (PLS-DA) is a supervised technique that is particularly useful in scoring the importance of significant variables in data. The important features are then used to discriminate variables for classification [14] or for dimensionality reduction [15]. PLS-DA assesses the influence of indi-

vidual features on the rest of components through calculating PLS coefficients which are called Variable Influence on Projection (VIP). Once the VIP scores are calculated, the model can score the importance of each variable based on a threshold value [16, 17].

PLS-DA was used in metabolomics for identifying metabolite that can be used as disease biomarkers. For example, PLS-DA was used for scoring the importance of variables which lead to identifying the biomarkers of Fulminant Hepatic Failure Disease (FHF) using the subsequent logistic regression which was used to predict the prognosis of FHF in mice [14]. PLS-DA was also used for finding the metabolic signatures of motor neuron disease [15].

2.2.2 SVM

Support Vector Machine (SVM) is a supervised learning technique that is used for both classification and regression. It is capable of finding robust solutions for classifications and regression applications. It has also a wide spectrum of complex problems while using much few parameters. SVM incorporates domain knowledge and uses kernel function to map the kernel space to a higher dimensionality by constructing and adjusting the linear boundaries of the support vector in the feature space. SVM was used in metabolomics in predicting genes functional classes and in studying drugs toxicity for classifying samples based on their metabolic biomarkers [18, 19].

2.2.3 MLP

Multilayer Perceptron (MLP) is an Artificial Neural Networks (ANN) algorithm based on mimicking human brain which consists of an interconnected set of nerve cells or neurons. The brain performs its function by using these neurons simultaneously [20, 21]. ANN depicts human brain by constructing a network of neurons that each can be viewed as elementary information processing unit [22]. It organizes these neurons into a set of hierarchical layers. Each layer consists of an input, output, and a hidden layer [23]. Figure 1 provides an illustration of MLP input, output, and hidden layers. Neurons are connected by links associated with numerical values that simulate long-term memory in the human brain by considering the strength and importance of each inputs. These values are called weights [17, 24].

The MLP back propagation algorithm uses Equations 1, 2 and 3 for updating weights.

$$\Delta W_j = \eta \delta_j O_i \tag{1}$$

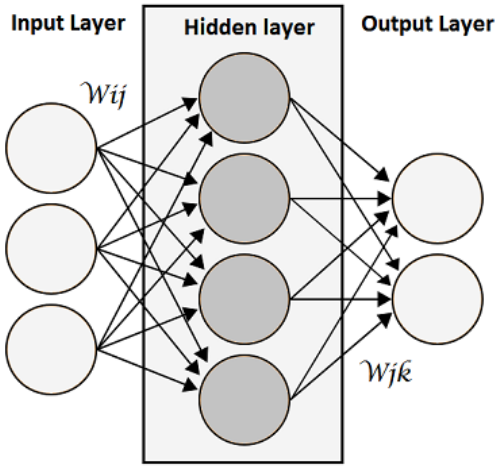

Figure 1. An illustration of the ANN layers.

$$\delta_j = O_j(1 - O_j)(T_j - O_j) \tag{2}$$

$$\delta_j = O_j(1 - O_j)\Sigma_K \delta_k W_{kj} \tag{3}$$

where W represents the neuron weight, η represents the learning rate and O_i represents the output calculated by neuron i for the output neurons. T_j represents the desired output for the neuron j in the internal hidden layer [25].

MLP is used for both supervised and unsupervised learning [23, 26]. It is used for predicting categorical values in sample classifications as well as for predicting continuous data in regressions [23]. In metabolomics, MLP was used medical applications, for predicting and tracking the severity of diseases based on quantitative metabolic data [27] and in combination with Principle Component Analysis (PCA) for identifying potential disease biomarkers [28].

2.2.4 Decision Trees (DT)

Decision Trees (DT) is a popular data mining technique [20, 22] that uses ID3 [29] and C4.5 [30] algorithms. Models built using decision trees are easy to interpret as they can be expressed as a set of logical rules and can be visualized using tree-like diagrams. Decision trees were successful in finding significant metabolites for diagnosis of breast cancer [31] and finding non-linear relationships in metabolomics data as well. In other applications, it was used for sample classification and ranking important metabolites in high dimensional data [27, 32].

2.3 Evaluation Methods

Five-fold cross-validation is a testing method that was used in this research in order to evaluate the validity of the results and assess the performance of

the created models in addition to its ability to be generalized using a separate dataset. In this method, the data is split into sub-datasets and the training (leave-one-out). The testing is repeated a number of times so that it covers all the created subsets. Each time, the model is trained using one sub-dataset while the remaining subsets are used for testing. The overall evaluation is calculated by combining the evolution results obtained for each sub-dataset.

The error rate was also used to compare the learning performance of the classifier in this study due to its relevance and simplicity. However, Classification Accuracy (CA), Sensitivity and Specificity were all used to evaluate the performance of the resulting data mining model. Equations 4,5,6, and 7 provide the mathematical foundation for these measures.

$$ErrorRate = 1 - CA \tag{4}$$

$$CA = (TP + TN)/(CP + CN) \tag{5}$$

$$Sensitivity = TP/CP \tag{6}$$

$$Specificity = TN/CN \tag{7}$$

where TP is the number of true-positive values and TN is the number of true-negative values. CP is the number of the actual positive values and CN is the number of the actual negative values [33].

Receiver Operator Characteristics (ROC) plot was also used as a performance measure. ROC is usually plotted using two axes. The x-axis represents false-positive rates while the second y-axis represents the true-positive rate [33, 34]. Each point on the curve represents the result of applying equation 8 and 9.

$$TruePositiveRate = TP/(TP + FN) \tag{8}$$

$$FalsePositiveRate = FP/(TP + TN) \tag{9}$$

where FN represents the number of false-negative values and FP represents the number of false-positive values [34]. TP and TN are as described above.

The Area Under the Receiver Operator Characteristics (AUC) was also used as a performance measure. It measures the performance of the machine learning model by calculating the area under the ROC curve.

3 Results

As the analytical objectives of the data-driven analysis performed in this study were set to: (1) developing a classifier for early prediction and building a tool to provide early diagnose of the disease and (2) identification of the significant metabolites that are more involved in the early stages of muscle loss. Since the bio-samples used in this study were of a urine origin, the chance contamination was quite high. Therefore, data preprocessing was needed in order to normalize the data using natural log transformation. Figure 2 on the next page shows the distribution of the data before and after the performed data normalization.

Data exploration was also needed in order to gain insight into the correlations between the targeted metabolites which helps to select the appropriate data mining technique that both suite the data and would fulfill the defined analytical objectives. The selection of these techniques was based on the results of data exploration and analytical objectives. Data exploration activities may include data investigation, data understanding, and data prospecting. The data has no missing values or significant outliers and therefore no additional preprocessing procedures were needed. Figure. 3 on the next page shows a heat map that was used in data exploration. It shows all possible correlations between all chemicals in the metabolite concentration profile.

Four data mining techniques were selected for the purpose of model building named as PLS-DA, MLP, SVM, and DT using ID3 and C4.5. Each of the models created using these techniques was evaluated using an independent iteration in order to provide the flexibility to forward the best model to the next phases. The performance evaluation of each model was conducted in terms of error rate. The selected techniques were found capable of achieving the analytical objectives while suiting the nature and quality of the dataset.

However, none of the applied data mining techniques performed well using the raw data. Therefore, a number of classical preprocessing procedures were conducted on the dataset in order to improve its quality and enhance the performance of the data mining models built using the selected techniques. The preprocessing procedures applied covered standardization, normalization and log transformation. Almost all the applied preprocessing procedures failed to improve the performance of the models with the exception log transformation. It slightly improved the performance of the MLP model. The performance of the created data mining models that were created using the raw and the preprocessed datasets is shown in Table 1 on page 84 using the classification error rate and evaluation measure that were discussed in Section 2.3.

Yet, when the techniques were applied to the log-transformed data, some improvement was noticed in the performance of SVM model along with some slight decline in the performance of the PLS-DA model. On the other hand, the best classification accuracy

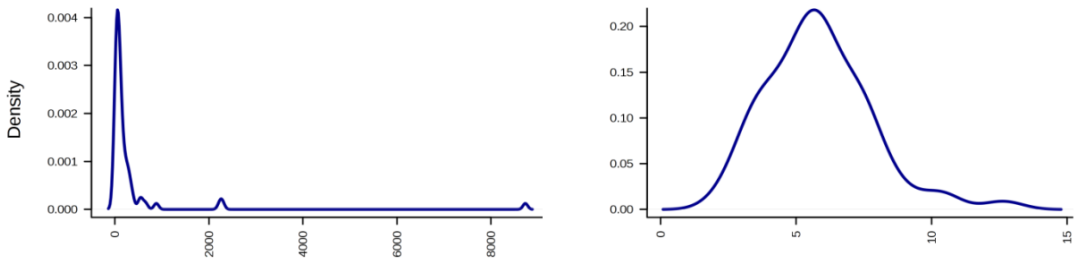

Figure 2. A density plot shows the data distribution before and after normalization

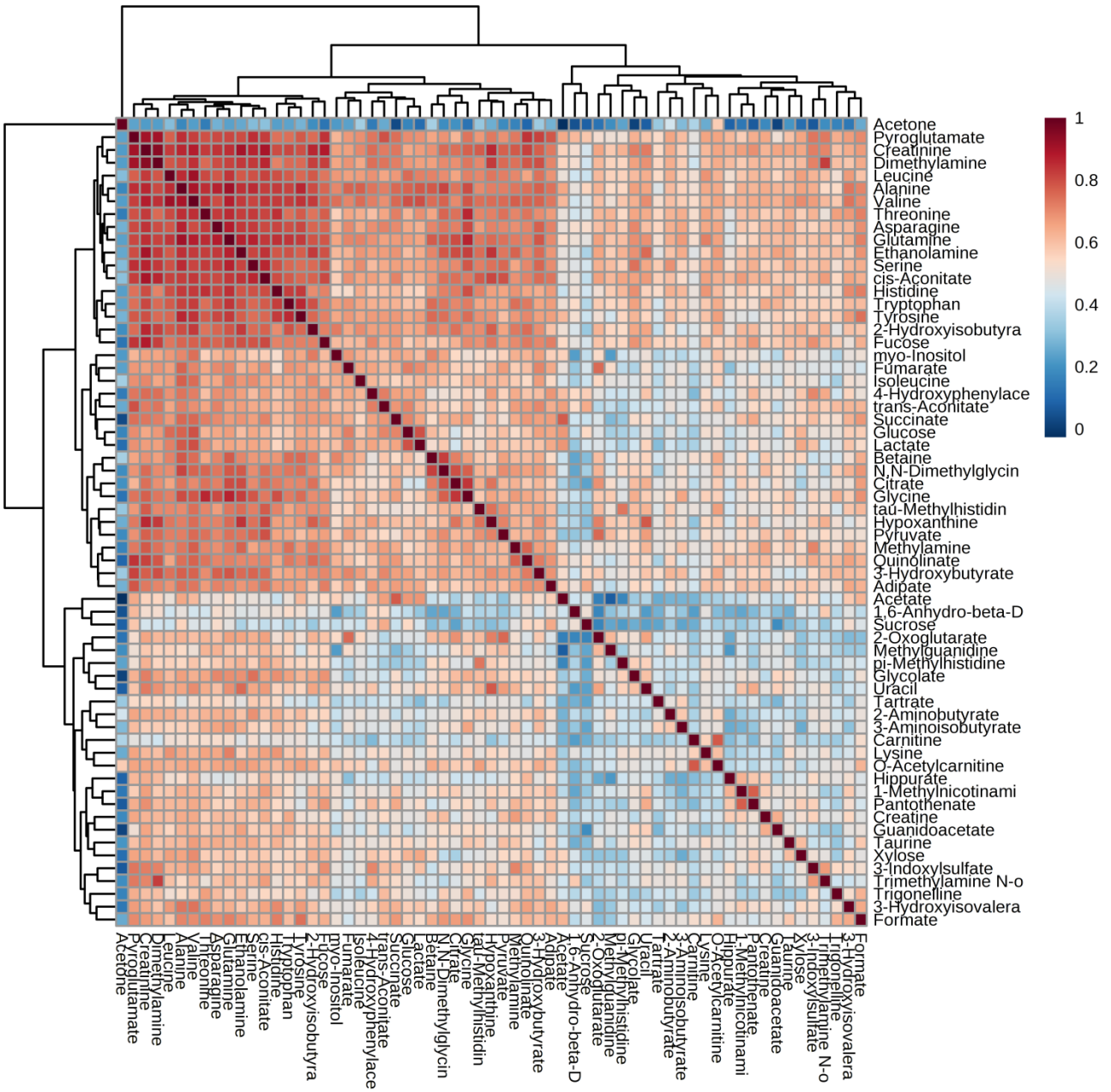

Figure 3. A heat map showing the correlation between metabolites: Dark colors show strong correlation, while light colors show weaker correlation.

Table 1. Classification error rate of the techniques over the original and the preprocessed complete dataset using: log transformation, standardization and normalization.

Log Transformation	*Standardization*	*Normalization*	*Classification Error Rate*				
			PLS-DA	*MLP*	*SVM*	*C4.5*	*ID3*
No	*None*	*None*	*30.1*	*34*	***27.3***	*33.8*	*39*
No	*Yes*	*None*	*30.1*	*32.2*	***27.3***	*33.8*	*39*
No	*None*	*Yes*	*34.6*	*40.5*	*34.3*	*39.7*	*39*
No	*Yes*	*Yes*	*34.6*	*35.1*	*34.3*	*39.7*	*39*
Yes	*None*	*None*	*30.9*	***27.3***	34.3	*33.8*	*39*
Yes	*Yes*	*None*	*32.2*	***26.5***	34.3	*33.8*	*39*
Yes	*None*	*Yes*	33	***26.5***	40.1	*37.1*	*39*
Yes	*Yes*	*Yes*	*34*	***26***	41	*37.1*	*39*

of most techniques was obtained when random re-sampling techniques were applied to the complete dataset.

Table 2 shows the performance of the models using random re-sampling compared to under-sampling and SMOTE over-sampling. However, despite the excellent classification accuracy achieved as shown in Table 2 which ranged between 89% and 91%, the random duplication of samples in the minority class may expose the model to the danger of over-fitting as it gives weight to some samples more than the others and may lead to false discovery.

Table 2. The classification error rate of the complete and the re-sampled dataset using SMOTE over-sampling, under-sampling and random re-sampling methods

Sampling	*Classification Error Rate*				
	PLS-DA	*MLP*	*SVM*	*C4.5*	*ID3*
None	*34.3*	*30.9*	*24.2*	*33.77*	*39*
SMOTE	*24*	*23.3*	*20.22*	*30.11*	*60*
Under-sampling	*35.6*	*32.9*	*27.9*	*33.21*	*63*
Random Re-sampling	***9.4***	***11.4***	***10.7***	***16.62***	*43*

Further data preprocessing procedures were decided in order to improve the results, which included data normalization: (1) Normalization by creatinine concentration [35], (2) Normalization by total peak area [36] and (3) Normalization by probability quotient [37]. However, none succeeded in improving the performance of the applied techniques. This is quite in line with the result of the earlier study.

More data exploration procedures were conducted in order to investigate the reduction of the classifier performance which led to identifying two issues with the data: (1) low ratio between the numbers of samples compared to the numbers of variables (2) the imbalanced distribution of the samples over the prediction classes. The number of samples in the cachexic class was 47, while in the control class was 30. The handling of the first issue involved reducing the number of variables to a narrower set of informative variables using IG, RF [38, 39] and SVM [40]. Two datasets were created for each data reduction method; one for the 30 most important variables and one for 13 most important metabolites (variables with significant SVM weight).

The sampling techniques were then applied to both: the complete and the reduced dataset as shown in Table 3 on the 85. The handling of the second issue involved applying three different re-sampling methods: Random Re-sampling, SMOTE Over-sampling and Under-sampling [13, 41]. Random resembling involves duplicating samples in the minority class in order to match the number of samples majority class. However, this may cause over-fitting particularly. Applying random re-sampling to the log-transformed dataset improved the performance of most techniques up to two folds except ID3 algorithm which performance declined slightly. Nevertheless, it was found that reducing the data to the most important 13 and 30 variables had only limited influence on the performance of the classifiers.

Under-sampling involves reducing the number of samples in the majority class so that the number of observation becomes equal to those in the minority classes which in fact increase the sensitivity of a classifier to the minority class. A decline in the performance of most techniques was observed when the techniques were applied to the under-sampled data. This decline was quite more severe with ID3 algorithm. However, the performance of PLS-DA and MLP improved somehow when these techniques were applied to the dataset that was reduced to 13 variables using SVM scorer. The selection of these variables was based on an assumed cutoff.

Table 3. The classification error rate of the models using the re-sampled and reduced datasets

Re-sampling Method	*Variables*	*Scorer*	*Classification Error Rate*				
			PLS-DA	*SVM*	*MLP*	*C4.5*	*ID3*
None	***13***	***IG***	*31.4*	*26.5*	*30.9*	*36.88*	*39*
None	***13***	***RF***	*28.6*	*26.5*	*29.1*	*34.81*	*39*
None	***13***	***SVM***	*31.4*	*26.5*	*30.7*	*36.88*	*39*
None	***30***	***IG***	*37.4*	*26*	*28.8*	*36.62*	*39*
None	***30***	***RF***	*35.3*	*23.4*	*29.9*	*35.58*	*39*
None	***30***	***SVM***	***20.3***	*24.2*	*20.3*	*34.55*	*39*
Random Re-sampling	***13***	***IG***	*26.8*	*22.3*	*22.3*	*18.96*	*43*
Random Re-sampling	***13***	***RF***	*22.6*	*22.6*	*22.6*	*18.7*	*43*
Random Re-sampling	***13***	***SVM***	*23.1*	*24.2*	*23.9*	*27.01*	*43*
Random Re-sampling	***30***	***IG***	*26.8*	*22.3*	*20.8*	*18.96*	*43*
Random Re-sampling	***30***	***RF***	*25.5*	*21.3*	*24.7*	*18.7*	*43*
Random Re-sampling	***30***	***SVM***	*24.2*	*21.3*	*21*	*16.62*	*43*
Under-sampling	***13***	***IG***	*32.5*	*28.9*	*30.7*	*34.64*	*62.9*
Under-sampling	***13***	***RF***	*36.1*	*27.5*	*32.2*	*27.5*	*62.9*
Under-sampling	***13***	***SVM***	***21.1***	*26.1*	***25.4***	*24.29*	*62.9*
Under-sampling	***30***	***IG***	*43.9*	*29.6*	*35.4*	*31.43*	*62.9*
Under-sampling	***30***	***RF***	*42.5*	*27.5*	*31*	*32.5*	*62.9*
Under-sampling	***30***	***SVM***	*30.7*	*27.5*	*28.2*	*30.71*	*62.9*
SMOTE	***13***	***IG***	*24.8*	*23.3*	*23.7*	*28.57*	*59.6*
SMOTE	***13***	***RF***	*27.7*	*24.2*	*27.3*	*29.45*	*59.6*
SMOTE	***13***	***SVM***	*25.3*	*25.9*	***14.9***	*28.57*	*59.6*
SMOTE	***30***	***IG***	*32.3*	*21.1*	*24.8*	*30.99*	*59.6*
SMOTE	***30***	***RF***	*29.5*	*22.4*	*23.5*	*29.89*	*59.6*
SMOTE	***30***	***SVM***	***32.3***	***21.1***	***27.3***	***30.99***	*59.6*

Over-sampling involves increasing the number of samples in the minority class to match their numbers in the majority class. Over-sampling was performed using SMOTE method [13] which involves the construction of a number of synthesized samples. SMOTE was successful in improving the performance classifiers prediction [10]. Over-sampling improved the performance of almost all techniques when the data was reduced to 13 and 30 variables apart from ID3 algorithm whose performance declined by one third and also when the data was reduced to 30 variables using SVM scorer. However, over-sampling had minor or no influence when the data was reduced using information gain and random forests techniques. Applying the rest of the data mining techniques to the reduced data decreased the performance of almost all the models regardless of the number of the reduced variables and the variable importance scorer used.

The confusion matrix in Table 4 shows a balanced distribution of true-positive and false-positive samples over both cachexia and control classes. Both classes scored the same accuracy. In addition, the ROC chart in Figure 4 demonstrated good model sensitivity as the ratio between true-positive and false-positive values was excellent. The model scored 92.6% in AUC performance measurement based on the ROC chart. Table 5 on page 87 shows the 30 important metabolites scored by SVM. Nearly half of the metabolites scored important in this research have been found consistent with those reported by the earlier study. These metabolites are underlined in the table. The SVM scorer agrees with 47% of the 30 important metabolites ranked by the bivariate analysis carried out in the earlier study. However, this percentage rises to 62% when considering the 21 important metabolites and fall back to 53% when it comes to the 13 most important metabolites.

4 Results and Discussion

Despite the fact that transformation, data reduction and sampling were successful in improving the performance of PLS-DA, SVM, MLP and C4.5, the best performance was achieved using PLS-DA model when applied to randomly re-sampled data as it scored 90.4% classification accuracy. However, this result was excluded as random res-sampling may expose the model to over-fitting by giving some samples more weight than others. The second best performance was achieved by MLP applied to over-sampled data using SMOTE. In addition to that most sampling techniques were successful in improving the performance of almost all classifiers except ID3, the best performance has been achieved using SMOTE method. On the other hand, normalization and standardization failed to improve the performance of any of the models, while log transformation made some slight improvement.

The performance of almost all classifiers improved when the dataset was reduced to the 30 and 13 most important variables using IG, RF and SV scorers except when ID3 was applied. Therefore, the SVM-MLP was selected as the champion and was then evaluated using five folds cross-validation. Half of the biomarkers were found identical to those reported in the earlier study, while the rest were also consistent with a lesser extent. In addition, the model was able to classify samples with 85.1% accuracy which outperforms all the models reported in the earlier study in almost every measure including: Error Rate, Classification Accuracy, Sensitivity, Specify, AUC and ROC.

In addition, the results reported in this research provide more insight and justification regarding the applied techniques. It also addresses the issue of imbalanced classes which was in fact, neglected in the original study. These improvements were achieved by applying a number of data transformation, preprocessing, sampling and data reduction procedures that were used in order to tackle the problem of imbalanced classes.

5 Conclusion

A data-driven analysis was carried out in this research focusing on analyzing the effect of preprocessing various procedures on the performance of the machine learning algorithms that were applied to a metabolite profile with imbalanced classes. An enhanced classifier was created in this research by combining SVM variable importance scoring with MLP classifier. The SVM-MLP model scored 85.2% classification accuracy and 85.1% in sensitivity and in recall measure which is an improvement when compared to the performance reported in the earlier study.

The study was successful in identifying a number of potential biomarkers for cachexia. The SVM-MLP model can be used clinically for developing an early clinical diagnosis method which can help to narrow down the laboratory analysis by targeting a smaller group of metabolites, which save time and reduced the diagnosis cost. The model is deployable using either in PMML or XML format and it can then be embedded in diagnosis software for the purpose of disease screening. The biomarkers identified can also be used for developing a heap and safe laboratory screening test before conducting the current more expensive and unsafe diagnosis procedures.

Table 4. The Confusion Matrix for the MLP Model

	Cachexic	Control	Total
Cachexic	***85.10%***	*14.90%*	*47*
Control	*14.90%*	***85.10%***	*47*
Total	*47*	*47*	***94***

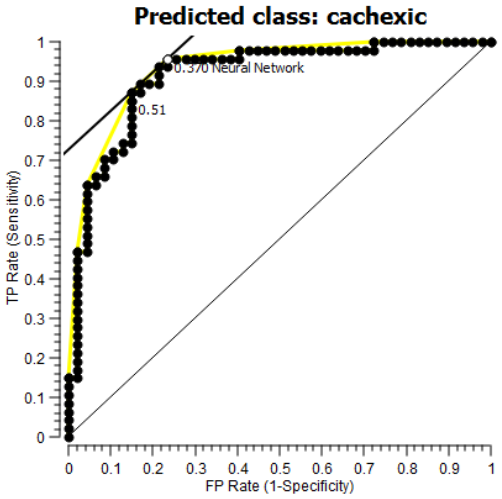

Figure 4. ROC Chart showing the relationship between true positive and false positive rate

Table 5. SVM Variable importance score using the re-sampled dataset. The underlined metabolites are those that were ranked important in this and the previous study

Variable Importance Score	*Metabolite*	*SVM Weight*
1	3-Aminoisobutyrate	*1.238*
2	Uracil	*1.163*
3	4-Hydroxyphenylacetate	*1.065*
4	**Glutamine**	*0.999*
5	**Glucose**	*0.98*
6	Pantothenate	*0.947*
7	Citrate	*0.873*
8	**Creatine**	*0.846*
9	**Leucine**	*0.818*
10	1-Methylnicotinamide	*0.802*
11	**Quinolinate**	*0.791*
12	**Glycine**	*0.764*
13	**Succinate**	*0.744*
14	Hypoxanthine	*0.679*
15	Isoleucine	*0.625*
16	**Tyrosine**	*0.612*
17	**myo-Inositol**	*0.522*
18	**Adipate**	*0.464*
19	**Methylamine**	*0.452*
20	**Trigonelline**	*0.407*
21	**Alanine**	*0.394*
22	2-Aminobutyrate	*0.329*
23	3-Hydroxybutyrate	*0.319*
24	3-Indoxylsulfate	*0.304*
25	**Acetate**	*0.302*
26	cis-Aconitate	*0.294*
27	Carnitine	*0.266*
28	Ethanolamine	*0.251*
29	1-6-Anhydro-beta-D-glucose	*0.251*
30	Lysine	*0.236*

References

[1] Roman Eisner, Cynthia Stretch, Thomas Eastman, Jianguo Xia, David Hau, Sambasivarao Damaraju, Russell Greiner, David S Wishart, and Vickie E Baracos. Learning to predict cancer-associated skeletal muscle wasting from 1h-nmr profiles of urinary metabolites. *Metabolomics*, 7(1):25–34, 2011.

[2] Vicki Maloney. Plant metabolomics. *BioTeach Journal*, 2:92–99, 2004.

[3] W. B. Dunn and D. I. Ellis. Metabolomics: Current analytical platforms and methodologies. *Trends in Analytical Chemistry*, 24(4):285–294, 2005.

[4] R. Goodacre, S. Vaidyanathan, W.B Dunn, G.G. Harrigan, and D.B. Kell. Metabolomics by numbers: Acquiring understanding global metabolite data. *Trends in Biotechnology*, 22(5):245–252, 2004.

[5] Ahmed Hmaidan Bani Mustafa and Nigel William Hardy. A strategy for selecting data mining techniques in metabolomics. In Nigel W. Hardy and Robert D. Hall, editors, *Plant Metabolomics: Methods and Protocols*, volume 860 of *Methods in Molecular Biology*, pages 317–335. Springer Science, 2012.

[6] A. BaniMustafa. Predicting software effort estimation using machine learning techniques. In *2018 8th International Conference on Computer Science and Information Technology (CSIT)*, volume 1, pages 249–256. IEEE, July 2018.

[7] R.J Bino, R.D Hall, O. Fiehn, and J. Kopka. Potential of metabolomics as a functional genomics tool. *Trends In Plant Science*, 9(9):418–425, 2004.

[8] Oliver Fiehn. Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. *Comparative and Functional Genomics*, 2:155–168, 2001.

[9] David S. Wishart. Metabolomics: applications to food science and nutrition research. *Trends in Food Science & Technology*, 19(9):482–493, 2008.

[10] Shaza M Abd Elrahman and Ajith Abraham. A review of class imbalance problem. *Journal of Network and Innovative Computing*, 1:332–340, 2013.

[11] Guang-Hui Fu, Feng Xu, Bing-Yang Zhang, and Lun-Zhao Yi. Stable variable selection of class-imbalanced data with precision-recall criterion. *Chemometrics and Intelligent Laboratory Systems*, 171:241–250, 2017.

[12] Sreejita Ghosh, E Baranowski, Rick van Veen, Gert-Jan de Vries, Michael Biehl, Wiebke Arlt, Peter Tino, and Kerstin Bunte. Comparison of strategies to learn from imbalanced classes for computer aided diagnosis of inborn steroidogenic disorders. In *Proc. of the European Symposium on Artificial Neural Networks*, 2017.

[13] NV Chawla, KW Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research (JAIR)*, 16:321–357, 2002.

[14] B. Feng, S. M. Wu, S. Lv, F. Liu, H. S. Chen, Y. Gao, F. T. Dong, and L. Wei. A novel scoring system for prognostic prediction in d-galactosamine/lipopolysaccharide-induced fulminant hepatic failure balb/c mice. *BMC Gastroenterol*, 9:99, 2009.

[15] Steve Rozen, Merit E. Cudkowicz, Mikhail Bogdanov, Wayne R. Matson, Bruce S. Kristal, Chris Beecher, Scott Harrison, Paul Vouros, Jimmy Flarakos, Karen Vigneau-Callahan, Theodore D. Matson, Kristyn M. Newhall, M. Flint Beal, Robert H. Brown, and Rima Kaddurah-Daouk. Metabolomic analysis and signatures in motor neuron disease. *Metabolomics*, 1(2):101–108, 2005.

[16] Anne-Laure Boulesteix and Korbinian Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1):32–44, 2007.

[17] Katherine Hollywood, Daniel R. Brison, and Royston Goodacre. Metabolomics: Current technologies and future trends. *Proteomics*, 6(17):4716–4723, 2006.

[18] SJ Barrett and WB Langdon. Advances in the application of machine learning techniques in drug discovery, design and development. In *Applications of Soft Computing*, pages 99–110. Springer, 2006.

[19] Young Truong, Xiaodong Lin, and Chris Beecher. Learning a complex metabolomic dataset using random forests and support vector machines. In Ronny Kohavi, Johannes Gehrke, William DuMouchel, and Joydeep Ghosh, editors, *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 835 – 840, Seattle, WA, 2004. ACM.

[20] Victor Maojo and José Sanandrés. A survey of data mining techniques. In G. Goos, J. Hartmanis, and J. van Leeuwen, editors, *Medical Data Analysis*, volume 1933 of *Lecture Notes in Computer Science*, pages 77–92. Springer Berlin / Heidelberg, 2000.

[21] Tom Mitchell. *Machine Learning*. McGraw-Hill series in computer science. McGraw-Hill, New York, 1997.

[22] Julien Boccard, Jean-Luc Veuthey, and Serge Rudaz. Knowledge discovery in metabolomics: An overview of ms data handling. *Journal of Separation Science*, 33(3):290–304, 2010.

[23] N. Jovanovic, V. Milutinovic, and Z. Obradovic. Foundations of predictive data mining. In *Neural Network Applications in Electrical Engineering, 2002. NEUREL '02. 2002 6th Seminar on*, pages 53–58, 2002.

[24] Michael Goebel and Le Gruenwald. A survey of data mining and knowledge discovery software tools. *SIGKDD Explor. Newsl.*, 1(1):20–33, 1999.

[25] S. Kotsiantis, I. Zaharakis, and P. Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, 2006.

[26] Ryszad S. Michalski, Ivan Bratko, and Miroslav Kubat. *Machine Learning and Data Mining: Methods and Applications*. John Wiley & Sons, Chichester, 1998.

[27] Royston Goodacre. Metabolomics of a superorganism. *Journal of Nutrition*, 137(1):259–266, 2007.

[28] Jin-mei Xia, Xiao-jian Wu, and Ying-jin Yuan. Integration of wavelet transform with pca and ann for metabolomics data-mining. *Metabolomics*, 3(4):531–537, 2007.

[29] R. Quinlan, J. Induction of decision trees. *Machine Learninig*, 1(1):81–106, 1986.

[30] R. Quinlan, J. *C4.5: Programming For Machine*

learning. Morgan Kufmann Publishing, USA, 1990.

[31] Jae Kim, Myoung Cho, Hyung Baek, Tae Ryu, Chang Yu, Myong Kim, Eiichiro Fukusaki, and Akio Kobayashi. Analysis of metabolite profile data using batch-learning self-organizing maps. *Journal of Plant Biology*, 50(4):517–521, 2007.

[32] David P. Enot, Manfred Beckmann, David Overy, and John Draper. Predicting interpretability of metabolome models based on behavior, putative identity, and biological relevance of explanatory signals. *Proceedings of the National Academy of Sciences*, 103(40):14865–14870, 2006.

[33] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[34] Paolo Sonego, András Kocsor, and Sándor Pongor. Roc analysis: applications to the classification of biological sequences and 3d structures. *Briefings in bioinformatics*, 9(3):198–209, 2008.

[35] Sushrut S. Waikar, Venkata S. Sabbisetti, and Joseph V. Bonventre. Normalization of urinary biomarkers to creatinine during changes in glomerular filtration rate. *Kidney Int*, 78(5):486–494, 2010.

[36] B. V. Stolyarov, A. G. Vitenberg, L. M. Kuznetsova, L. N. Ogongo, and S. A. Smirnova. A modification of the internal normalization method with sample splitting. *Chromatographia*, 9(1):3–9, 1976.

[37] Frank Dieterle, Alfred Ross, Götz Schlotterbeck, and Hans Senn. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in 1h nmr metabonomics. *Analytical Chemistry*, 78(13):4281–4290, 2006.

[38] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001.

[39] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–58, 2003.

[40] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, Pittsburgh, Pennsylvania, United States, 1992. ACM.

[41] Anne H. Milley, James D. Seabolt, and John S. Williams. Data mining and the case for sampling solving business problems. Technical report, SAS Institute Inc, 1998.

Ahmed BaniMustafa is an assistant professor of Computer Science at the American University of Madaba. His major research interest is Data Mining, Bioinformatics, Software Engineering, Data Science, and Machine Learning. He received his PhD in Computer Science from Aberystwyth University in United Kingdom in 2012 and his MSc in Software Engineering from UWE Bristol University. Dr. Ahmed has joined the American University of Madaba in 2013 as one of the founders of Computer Science Department where he created an undergraduate program in Data Science. He has supervised a number of innovation and dissertation projects in this department. He worked earlier for the Ministry of Education, Philadelphia University, Jordan University of Science & Technology, Aberystwyth University and Motorola Mobility which was acquired later by Google Inc. Dr. Ahmed is a member of ACM, IEEE, JSSR, and Metabolomics Society.